Collective Innovation in Groups of Large Language Models

Eleni Nisioti, Sebastian Risi, Ida Momennejad, Pierre-Yves Oudeyer, Clément Moulin-Frier

Collective Innovation in Groups of Large Language Models

Eleni Nisioti[1], Sebastian Risi [1], Ida Momennejad [2], Pierre-Yves Oudeyer [3] and Clément Moulin-Frier [3]
[1] IT University, Denmark
[2] Microsoft Research, United States
[3] Inria Center of the University of Bordeaux, France
enis@itu.dk

Abstract

Human culture relies on collective innovation: our ability to continuously explore how existing elements in our environment can be combined to create new ones. Language is hypothesized to play a key role in human culture, driving individual cognitive capacities and shaping communication. Yet the majority of models of collective innovation assign no cognitive capacities or language abilities to agents. Here, we contribute a computational study of collective innovation where agents are Large Language Models (LLMs) that play Little Alchemy 2, a creative video game originally developed for humans that, as we argue, captures useful aspects of innovation landscapes not present in previous test-beds. We, first, study an LLM in isolation and discover that it exhibits both useful skills and crucial limitations. We, then, study groups of LLMs that share information related to their behaviour and focus on the effect of social connectivity on collective performance. In agreement with previous human and computational studies, we observe that groups with dynamic connectivity out-compete fully-connected groups. Our work reveals opportunities and challenges for future studies of collective innovation that are becoming increasingly relevant as Generative Artificial Intelligence algorithms and humans innovate alongside each other.

Introduction

Human culture evolves through the accumulation of artefacts, semantic repertoires, and behaviours that become more complex over long time-scales (Creanza et al., 2017; Solé et al., 2013; Whiten et al., 2021a). While popular narratives often emphasize the contributions of lonely geniuses (Montuori and Purser, 1995), a different story emerges if we consider historical data on cultural artefacts (Solé et al., 2013; Eldredge, 2011; Pereira et al., 2023), theoretical analysis (Creanza et al., 2017; Solé et al., 2013), human behavioural experiments (Derex and Boyd, 2016; Mason et al., 2008; Brackbill and Centola, 2020) and computational studies (Lazer and Friedman, 2007; Cantor et al., 2021; Brackbill, 2017; Fang et al., 2010). This body of work suggests that human cultural evolution is an inherently collective process, akin to biological evolution (Creanza et al., 2017; Solé et al., 2013): innovations arise in a collective as individuals modify and recombine existing ones in their environment. In this work, we contribute novel computational evidence to support this hypothesis. We employ groups of Large Language Models (LLMs) to solve innovation tasks, examining how they perform in isolation and how their social connectivity affects their collective behaviour.

While many species exhibit culture (Whiten et al., 1999; Aplin, 2019), cultural change (Aplin et al., 2015), and even a continuously complexifying cultural repertoire (Whiten et al., 2021b), humans are unique in their ability to accumulate innovations (Tennie et al., 2009; Derex, 2021; Boyd and Richerson, 1996). Studies aiming at understanding this phenomenon have drawn links between collective innovation and, among others, social learning capacities (Dunbar, 1993a; Lotem et al., 2017), group size (Dunbar, 1993b; Kline and Boyd, 2010; Derex et al., 2013) and social connectivity (Lazer and Friedman, 2007; Cantor et al., 2021; Brackbill, 2017; Fang et al., 2010; Nisioti et al., 2022).

Despite a plurality of studies and methodologies, recent positions in cultural evolution (CE) warn that the field may have neglected certain mechanisms, such as the role of individual cognition in innovation and invention (Singh et al., 2021; Perry et al., 2021; Smolla et al., 2021). The majority of collective innovation studies model individual inventions solely as random mutations and recombinations of existing innovations (Mason et al., 2008; Cantor et al., 2021; Fang et al., 2010). Yet a number of hypotheses in human studies point to the important role that language and, in general, advanced cognitive mechanisms, have played in our evolution (Dunbar, 1993b; Boyd, 2018; Dunbar, 2014; Smith et al., 2017). Equipping models of CE with language can enable the study of such hypotheses and reduce the artificiality of experimental set-ups by bringing them closer to the ones used with human participants (Mesoudi, 2021).

Beyond their capacity to model human language, LLMs have shown an emergent ability to model certain aspects of human behaviour, such as fairness in economical decisions (Horton, 2023), content biases in information transmission (Acerbi and Stubbersfield, 2023), and convincing social interactions in realistic social simulation games (Park

Figure 1: Studying collective innovation in groups of LLMs: A) we experiment with Little Alchemy 2 (LA2), a game where players combine real-world items to create new ones. A knowledge graph describes the possible combinations (we only present a small sub-part of the graph which contains 720 items in total) B) Alice-LLM and Bob-LLM are two LLMs playing the game together. They are provided with the same intro prompt, explaining the rules of the game, and the same task (they start with the same set of items). Alice-LLM and Bob-LLM have identical weights but behave differently because the state prompt depends on their crafting history. They are informed about the actions of others through their prompt. In this paper, we study how groups of such LLM agents are able to efficiently explore a knowledge graph, focusing in particular on the effect of different social structures specifying with whom and when they can share information

et al., 2023). Moreover, LLMs are evermore present as copilots of human labour and have become actors in the process human cultural evolution (Brinkmann et al., 2023). A natural next question is whether they can also be useful in computational studies of cultural evolution as generative models of individuals (Perez et al., 2024).

Here, we study how groups of LLMs solve collective innovation tasks. As a test-bed, we employ Little Alchemy 2 (LA2) [1], an existing creative game where players need to combine items to create new ones. To identify which combinations are valid, Little Alchemy 2 employs a knowledge graph where items are real-world entities and combinations are inspired by our physical reality (for example, 'fire' and 'water' results in a new item, 'steam'). Little Alchemy 2 was recently proposed as a test-bed for the study of human exploration, as it poses challenges not present in classically-employed bandit tasks (Brändle et al., 2023). Here, we test a similar proposal, namely that Little Alchemy 2 can be a useful test-best for studying both human and computational cultural evolution. While Little Alchemy 2's knowledge graph is certainly not a comprehensive database of human cultural artefacts, it is significantly more realistic than previous test-beds containing a few hand-crafted interactions among symbolic items (Derex and Boyd, 2016; Cantor et al., 2021; Migliano et al., 2020). We present a visualization of a small part of the knowledge graph of LA2 in Figure 1.

First, we examine an LLM in isolation to probe its problem-solving capacities independently of group influences. We identify three key challenges related to a) *factual knowledge*: to efficiently explore the search space an LLM needs to leverage semantic knowledge about the task, b) *multi-step reasoning*: to reach a target item a player often needs to craft multiple intermediate items, and c) *exploration*: a key feature of LA2 is its open-ended nature. As is common in creative games, no instructions are given to the player, who starts with a couple of items and is left to explore a vast search space. As previous studies with human participants have shown, non-random exploration strategies, such as empowerment, are required to explore efficiently (Brändle et al., 2023; Klyubin et al., 2005). To our knowledge, the abilities of LLMs to exhibit such forms of exploration have not been studied before.

To examine the knowledge and multistep reasoning abilities of LLMs we first study innovation tasks that require crafting a target item. These tasks were originally introduced by Jiang et al. (2020), and their complexity can be controlled by determining the number of intermediate items the player needs to craft before reaching the target, termed their depth. Our experiments in tasks with a target suggest that: a) LLMs leverage factual/ knowledge, as removing the natural language semantics degrades performance b) multi-step reasoning is challenging as performance significantly drops when increasing the depth of the task.

We, then, focus on tasks without a target that employ the same graph and starting set of items with LA2 . We, first, examine the ability of a single LLM to explore the search space and compare it with a baseline agent that employs empowerment and is known to perform on par with humans (Brändle et al., 2023; Klyubin et al., 2005). Our experiments indicate that the LLM agent performs on par with an agent that randomly chooses combinations, suggesting that it does not explore efficiently. Following the single-agent LLM experi-

[1] https://littlealchemy2.com

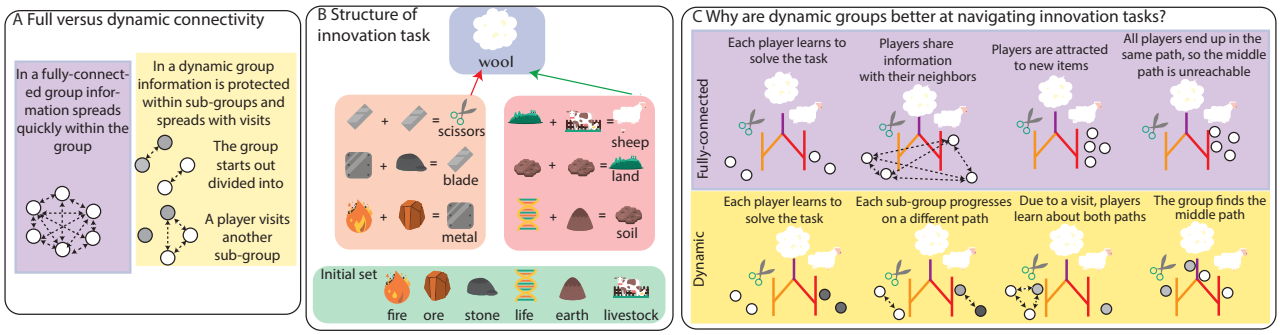

Figure 2: Examining the effect of social connectivity on collective innovation: A) We consider two types of connectivity: a fully-connected group of 6 agents and a group with dynamic connectivity that starts out with agents divided into 3 sub-groups of two agents and visits of a fixed duration take place between groups with a random probability. B) Example structure of an innovation task in LA2 : the task starts out with 6 items that the player can combine to move up the search space. Depending on their semantics the items may create independent trajectories as the two presented here, culminating into the items "scissors" and "sheep". To discover the item "wool" the player needs to reach the end of both trajectories and combine them. C) Why are dynamically-connected groups better at solving this innovation task? When a player observes that another player has found a new item it follows along. This means that, in a fully-connected group, all agents will get trapped in a single path. In contrast, subgroups of a dynamic group may explore different paths and, then, manage to recombine their solutions.

ments, we study the same tasks in a multi-LLM setting with 6 LLMs solving the same task. As we show on the right of Figure 1, LLMs in a group share information solely through the prompt: similarly to previous human lab studies (Derex and Boyd, 2016; Migliano et al., 2020), the history of crafting actions of its neighbors is made available to the LLM. To study the effect of social structure, we compare the performance of a fully-connected group, where all agents communicate and share information with all other agents, and a group with dynamic connectivity employed in previous studies (Derex and Boyd, 2016; Nisioti et al., 2022; Cantor et al., 2021) (we illustrate the two types of connectivity on the left of Figure 2). In our multi-agent experiments, we observe that: a) LLMs learn imperfectly from social information, as there is some delay between the moment a neighbour of the LLM crafts an item and the LLM crafts it itself. We refer to this phenomenon as "imperfect copying" b) provided that they perfectly copy their neighbors, LLMs perform better in a collective than in isolation c) under the same assumption, groups with dynamic connectivity out-compete fully-connected groups. This observation agrees with previous human and computational studies, which hypothesize that partially-connected groups are at an advantage due to the tree-like structure of innovation landscapes (Derex and Boyd, 2016; Stanley and Lehman, 2015) (we illustrate this hypothesis on the right of Figure 2).

Related work

Various computational studies have investigated the dynamics of collective innovation. These studies differ across three important dimensions. First, they may model individual and social learning mechanisms differently. The majority of CE models assumes that innovations arise solely through random mutations and recombinations (Mason et al., 2008; Cantor et al., 2021; Fang et al., 2010), nullifying the cognitive capacities of individuals. Considering the problem-solving abilities of LLMs, discussed right after, our model can be seen as being rather closer to works that equip agents with cognition, for example through reinforcement learning (Nisioti et al., 2022). Second, studies propose different mechanisms as drivers of cultural accumulation. Here, works may focus on the effect of social connectivity, with some suggesting that more connectivity is better and size is the sole determinant of collective performance (Mason and Watts, 2012), while others suggesting that partial, static or dynamic, connectivity confers an advantage to groups exploring search spaces with local optima (Derex and Boyd, 2016; Nisioti et al., 2022; Lazer and Friedman, 2007) Third, they employ different test-beds to capture the landscape of innovation. Initially computational studies studied classical search problems such as line search (Mason et al., 2008; Mason and Watts, 2012) and the NK-problem (Lazer and Friedman, 2007), thus ignoring the hierarchical, tree-like structure of innovation landscapes. Another line of works employs a test-bed inspired from drug discovery, where individuals need to combine items to craft new ones (Derex and Boyd, 2016; Cantor et al., 2021; Nisioti et al., 2022), and is, thus, closest to the test-bed employed here. Yet this test-bed differs from ours in two ways: a) it is defined over symbolic items and manually designed small-scale knowledge graphs, thus failing to capture the scale and semantics of human innovation b) it assumes that each combination incurs a reward for the agent. In contrast, our open-ended tasks instruct the agent to explore and may, thus, capture the open-ended na-

ture of human innovation (Fogarty et al., 2015).

The cognitive capacities of LLMs have been under close scrutiny, with studies revealing surprising limitations considering their convincing conversational skills (Mitchell and Krakauer, 2023). Such studies have focused on their factuality, planning and reasoning abilities, showing that LLMs are prone to hallucinations and falling into self-reinforcing loops (Momennejad et al., 2023; Wei et al., 2022). Multi-agent LLM studies (Webb et al., 2024; Zhuge et al., 2023) have revealed interesting phenomena reminiscent of collective intelligence with LLMs, such as that having them debate in a group prior to solving a task improves both their factuality and reasoning abilities (Du et al., 2023a; Zhuge et al., 2023). An understudied aspect of LLM cognition is, however, their ability to explore. While LLMs have been employed for generating and evaluating goals for RL agents (Colas et al., 2023; Du et al., 2023b) and exploring in open-ended tasks (Wang et al., 2023a), such studies did not examine whether LLMs improved exploration due to leveraging semantic knowledge or planning. When LLMs were evaluated on their ability to learn how to explore arbitrary bandits tasks solely through in-context learning, experiments indicated that only the state-of-the-art GPT-4 model, equipped with external memorization mechanisms, successfully engages in exploration (Krishnamurthy et al., 2024).

Collective innovation test-bed for LLMs

We, here, describe the test-bed we have designed to study collective innovation with LLMs. We have minimally extended Wordcraft, a Python-based gym reinforcement learing (RL) environment inspired by Little Alchemy 2 that was originally introduced for evaluating the commonsense abilities of RL agents (Jiang et al., 2020). Instead of images, items in Wordcraft are expressed as text. Another difference with the original game is that tasks are targeted: whereas a game in Little Alchemy 2 starts with the player being given a small set of items and no further guidance, tasks in Wordcraft determine, in addition to the initial set, a target item to craft. Our extension of Wordcraft is basically an introduction of open-ended tasks that follow the spirit of Little Alchemy 2 and an interface for mapping the gym environment to a textual form. To capture the diversity of the task design space of our test-bed, below we provide a general definition of its components and explain our particular implementation. These components are:

A knowledge graph that represents the semantics of the task space by indicating how items can be combined to create new items. In this work we employ the knowledge graph of Little Alchemy 2, but, in the general case, studies can manually define their own graphs (Nisioti et al., 2022) or derive graphs from text corpora (Jiang et al., 2020).

A task-generation process that samples the initial set of items from the knowledge graph and determines the goal of the task. Here, we discriminate between open-ended and targeted tasks. The former prompt the LLM to discover as many items as possible without specifying a target. The latter prompt the LLM to craft a specific item and their complexity can be configured through two parameters: the number of items inserted in the initial set that are irrelevant for crafting the current target, w, termed *distractors*, and the number of intermediate items that the LLM needs to craft before reaching the target, termed the *depth*.

A textual repesentation of the task that has two parts: a) an intro prompt containing the rules of the game, as well as examples of tasks and correct outputs from the LLM in order to elicit in-context learning (Brown et al., 2020) b) the current task state. As a task requires multiple crafting steps (in the case of open-ended tasks on the scale of hundreds), we need a way to keep the prompt size limited. For this reason, instead of showing the complete crafting/discussion history for a task we summarize its current state in the following form for targeted tasks:

<Current task>
Inventory: set of items available for crafting
Target: target item
Remaining rounds: number of crafting steps before task is over
Task valid combinations: a list of combinations already attempted that gave a new item
Task invalid combinations: a list of combinations already attempted that did not give a new item

In the case of open-ended tasks, information about the target is omitted. In addition to keeping the prompt size limited, this way of presenting the task can be seen as a form of external summarization (Krishnamurthy et al., 2024), as the LLM does not need to memorize the item combinations it has attempted. We provide an illustration of intro and task prompts in an example task where two LLMs are solving an open-ended task collectively on the right of Figure 1.

More formally, a task in our test-bed is described by a set of items that are initially available for crafting, $\mathcal{I}_0$, a knowledge graph $\mathcal{K}$ and a target item g (that is empty in the case of open-ended tasks). The initial set $\mathcal{I}$ and target g are sampled from the knowledge graph $\mathcal{K}$ at the beginning of an episode that lasts for a fixed number of steps T. The agent can execute actions in the form of two items, $[i_1, i_2]$ that denote the combination it attempts to make. If, according to the knowledge graph, the action is a valid combination the environment returns a new item that is inserted in the inventory of the agent (if the item is already it is discarded, as items can be re-used indefinitely). At each step t, the agent is presented with the current state of the environment and is prompted to execute an action.

Baselines

Here we describe previous models that do not make use of LLMs and have been employed in single-agent and collec-

tive innovation studies.

Singe-agent

Empowered agent Empowerment is a type of exploration strategy that centers around the idea of choosing actions that enable the generation of as many options as possible in the future (Klyubin et al., 2005). Previous lab studies where human participants played LA2 have shown that empowerment is a powerful exploration strategy in this game and that it captures human behavior more accurately than other exploration strategies, such as random exploration and uncertainty-minimization (Brändle et al., 2023). While different implementations of empowerment are possible, here we consider the one employed in (Brändle et al., 2023): at each timestep t an agent employing empowerment combines the two items in its inventory that will result in crafting the most empowering item. The empowerment of an item is computed as the number of valid combinations it participates in. To choose an action, this agent accesses the knowledge graph of the task (defined in the previous section), computes the empowerment value of each potential combination and chooses the one with the highest value.

Random agent At each timestep t, this agent randomly combines two items from its inventory.

Multi-agent

Random groups The majority of computational studies in collective innovation (Mason et al., 2008; Cantor et al., 2021; Fang et al., 2010; Nisioti et al., 2022) consider random agents (as defined in the previous paragraph) that can combine items they see in the inventories of their neighbours. This implies that they have a perfect mechanism for copying social information. Agents can introduce new items by randomly mutating a single item or combining multiple items. For brevity, we refer to such groups of agents as random groups.

Empowered groups Here we have multiple empowered agents that can combine items they see in the inventories of their neighbors.

Collective of LLMs with social connectivity

Here, we describe how we designed group sof LLM agents for solving tasks in our innovation test-bed. We first describe a single LLM agent and, then, describe how they share information in a group with a certain social connectivity.

At each timestep t an LLM agent outputs text that contains the action it chooses to execute based on the prompt describing the task (described in the previous section). The prompt has been engineered to instruct the LLM to provide its output in the specific format:

> *Combination*: 'first item' and 'second item'
> *Reasoning*: based on the information in the <Current task>, do reasoning about why you chose this combination

Instructing the LLM to follow a certain format is a common practice when it needs to interface with another system, such as an RL environment (Wang et al., 2023b). Prompting the LLM to reason on its response is also a common technique, termed chain-of-thought prompting, and has been shown to improve the reasoning abilities of LLMs (Wei et al., 2022). As the LLM output probabilities over tokens, we can control the randomness of its outputs through the temperature of a soft-max function over probabilities.

We consider groups of identical LLMs (they all have identical weights) that differ solely in the prompt that describes the task. As we showed in Figure 1, the different LLMs are presented with the same intro prompt and are also assigned with the same task. What differentiates LLMs is the state prompt, i.e., the current state of their inventory and their history of past combinations. Thus, any differences in the outputs of LLMs in a given group solely arise due to their own behavior in the current task and are bootstrapped by the randomness in their sampling strategies.

A group is characterized by its social connectivity, an undirected graph G that determines the local neighbourhood of an agent. In this work, we consider two types of connectivity: a) in fully-connected groups all LLM agents are neighbours with each other b) in dynamic groups the group is divided into sub-groups of two and agents visit another random sub-group with probability p for a fixed duration V (see left of Figure 2 for an illustration of the two connectivities).

Agents in a group interact solely by sharing information regarding their actions. In particular, social information is shared by augmenting the prompt with the following tags:

> *Other players' valid combinations*: a list of combinations already attempted by other players in the agent's neighborhood that gave a new item
> *Other players' invalid combinations*: a list of combinations already attempted by other players in the agent's neighborhood that did not give a new item

Thus, similarly to previous computational studies (Mason et al., 2008; Cantor et al., 2021; Fang et al., 2010; Nisioti et al., 2022), information is shared implicitly based on the social connectivity without requiring an action on behalf of the agents. We illustrate how socially-shared information is presented to the agents on the right of Figure 1 and the two types of social connectivity that we consider in our study on the left of Figure 2.

Results

Set-up

We perform an experimental analysis where we control for different features of our set-up: a) we consider both targeted and open-ended tasks. For the former we randomly sample tasks with different number of distractors ($w \in \{3, 6\}$) and different depth values ($d \in \{1, 2\}$). We allow six crafting steps and measure success as the percentage of tasks where the agent crafts the target item. We perform 10 trials and, for each one, we randomly sample 50 tasks. Tasks are identical across trial. Thus, variance within a trial reveals the effect of task variability while variance across trials reveals variability in the agent's policy for the same task. For the open-ended tasks we employ the same initial set with the one used in Little Alchemy 2 ('air', 'earth','fire','water'), allow 200 crafting steps and average across 10 trials. Here, we employ the size of the agent's inventory as a proxy for success b) we study two different LLMs, GPT-3.5 turbo [2] and an open-source model, Llama 2 [3] and compare their performance to the two single-agent baselines (random agents and agents using empowerment with a temperature of 0.1). We employ a temperature value of 1.0 for the LLMs (we performed a grid search for Llama 2 and did not observe large variations but expect that a high temperature value is useful for eliciting randomness across agents) c) we study groups with two types of social connectivity: a fully-connected group with 6 agents and a dynamic group where agents are divided into subgroups of two and, unless otherwise specified, perform visits with a probability of $p = 0.2$ that last for V =50 steps. To avoid repetitions, if an agent repeats an already-attempted combination, we re-prompt until it chooses a novel one or 6 crafting steps have passed. We provide code for reproducing experiments, including prompts, in an online repo.

LLMs can exploit task knowledge

To gauge the ability of LLMs to understand and solve innovation tasks, we first examine the percentage of solved targeted tasks, S, for varying task complexity with single-agent methods in Figure 3. Error bars indicate variance due to both task and agent variability. We observe that when $w = 3, d = 1$ GPT-3.5 turbo solves almost all tasks ($S = 0.9 \pm 0.1$) while other methods succeed about half of the time. As these tasks contain 5 items, they can be solved successfully by exhaustive search within 10 steps and, by random search, about half of the time within the time budget. Agents employing empowerment outperform random ones, as empowering items have higher chances of being targets. Llama 2 performs significantly worse than GPT-3.5 turbo , an observation that generalizes to all settings we examined. A major reason for this is the inability of Llama 2 to avoid repeating combinations. In particular, we counted the number of times an LLM repeats a combination and observed that it was close to zero for GPT-3.5 turbo but Llama 2 repeats itself an overage of 4 times per task and in many cases, never finds a novel combination despite the repetition mechanism.

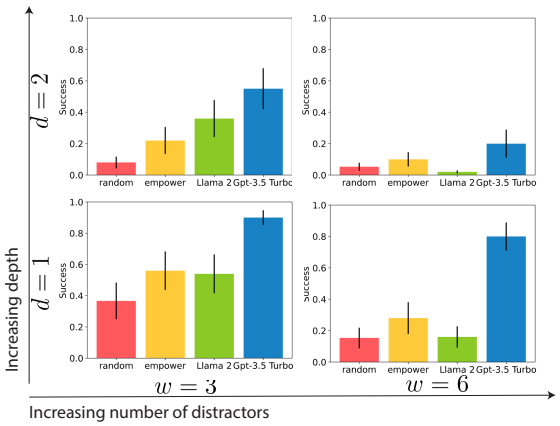

Figure 3: Performance in targeted tasks of varying complexity: success is the percentage of tasks for which the agent crafted the target item within the allowed time budget.

To understand these differences in performances, we perform two additional probing tasks that aim at examining the knowledge of LLMs. First, we test to what extent the performance of LLMs is influenced by the semantics of the task. We do so by encoding all words in the knowledge graph into random strings of 5 characters. We present how success changes for the two LLMs when removing semantics in Table 4, where we observe that the performance of both LLMs drops when semantics are removed. The drop is much steeper for GPT-3.5 turbo , which is not surprising, as it performs much better than Llama 2 at the tasks with semantics. The fact that semantics are crucial when there is one level indicates that the LLMs can predict which combination is the most likely to give the target item in a single step.

A yet more challenging and particularly useful skill, considering the multi-level nature of tasks, is predicting the outcome of crafting. To test for this, we create another probing task: we prompt the LLM with valid combinations (generated by sampling combinations with the agent employing empowerment) and ask the LLM to predict the outcome of the combination. We, then, compute the similarity between these predictions with the actual crafting outcome using the pre-trained glove model 'glove-twitter-25' [4]. We present these results on the right of Figure 4, where we include a baseline that randomly samples items from the knowledge graph as predictions for reference (similarity values range between 0 and 1). We observe that GPT-3.5 turbo is better

[2] https://github.com/meta-llama/llama

[3] We use the 13-billion parameter model from https://chat.openai.com/auth/login

[4] https://huggingface.co/Gensim/glove-twitter-25

Success

	With semantics	Without semantics
Llama 2	0.54	0.23
GPT 3.5	0.9	0.5

Prediction accuracy of crafted item

Llama 2	0.59
GPT 3.5	0.67
random	0.51

Figure 4: Examining the knowledge of LLMs in two probing tasks: (a) effect of removing semantics form the knowledge graph (b) semantic similarity (computed using glove embeddings) between the crafting outcomes predicted by the LLMs and the actual outcomes

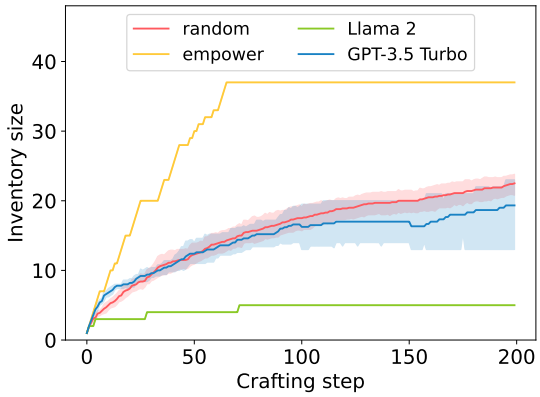

Figure 5: Performance in open-ended tasks for single agents

than Llama 2 at predicting crafting outcomes but is far from perfect (the online repo includes details about this probing task, including the prompt and predictions of the LLMs).

LLMs struggle at multi-step reasoning

As we increase the complexity of the tasks we observe, in Figure 3 that: a) when the number of distractors increases ($w = 6, d = 1$) the gap between GPT-3.5 turbo and other methods increases. This is not surprising as GPT-3.5 turbo exploits task knowledge best and the benefits of informed exploration become more apparent in larger search spaces b) when the depth increases ($w = 3, d = 2$) then the performance drop is more significant and the disparity between the two LLM models decreases. Combined with the observation that GPT-3.5 turbo cannot perfectly predict outcomes this suggests that multi-step search poses a qualitatively different challenge that even advanced LLMs cannot solve.

LLMs struggle in open-ended tasks

We move to the open-ended task, where we observe that: a) Llama 2 only crafts an average of 6 items, performing worst among all methods. As discussed earlier, Llama 2 has the tendency to repeat combinations which is particularly detrimental in open-ended tasks due to the long horizon without

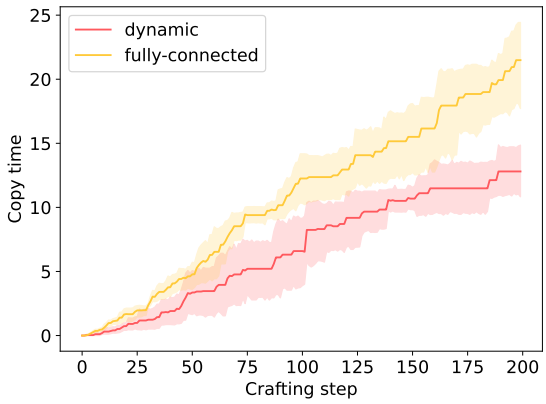

Figure 6: Copy time (number of timesteps it takes for a valid combination to appear in the inventory of an agent once it has appeared in one of its neighbors) for multi-agent GPT-3.5 turbo

resets b) GPT-3.5 turbo performs on par with random. This suggests that it does not leverage its knowledge for efficient exploration. This model also repeats combinations but at a much smaller rate. c) empowered agents perform best. This is not surprising as empowered agents have access to the knowledge graph and employ an exploration objective that is useful in such multi-level tasks (Brändle et al., 2023).

LLMs copy others imperfectly

We now move to experiments with groups where we first examine the collective behavior of GPT-3.5 turbo agents (we do not experiment with Llama 2 groups due to their suboptimal performance). A first question is whether LLM-agents benefit from innovating in a collective and, a requirement for this, is learning from the combinations of others (we have described how social information is shared when presenting the methods). A proxy for this is the tendency of agents to copy actions they see in others. To search for this we count the number of crafting steps it takes for an element to appear in the inventory of an agent once it appears in the inventory of one of its neighbors. We compare the evolution of this variable for dynamic and fully-connected groups in Figure 6 where we average within groups. We observe that copying is not perfect: as time passes more and more items accumulate that the agent could have crafted. Agents in fully-connected groups take more time to copy their neighbors. Potential reasons for this are that copying when having more neighbors takes more time, as more items need to be copied, and that larger neighborhoods lead to longer prompts, making the task more difficult for LLMs.

Connectivity influences collective innovation

Finally, we compare groups with different connectivity in terms of their performance in open-ended tasks. Figure 7

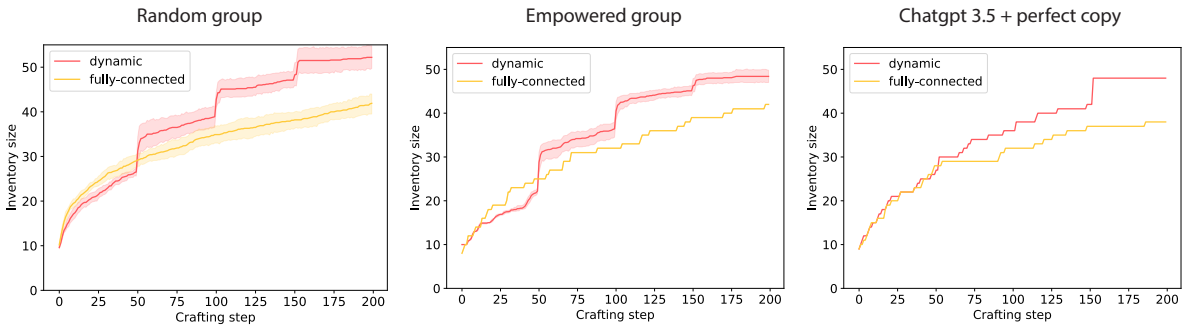

Figure 7: Effect of social connectivity in open-ended tasks with different agent models

reveals that, for all groups, dynamic connectivity performs best. In the case of GPT-3.5 turbo we present results with a perfect copy mechanism (when an agents crafts a new item, it appears in the inventories of all its neighbors). As we saw in the previous paragraph, GPT-3.5 turbo agents do not copy perfectly and are therefore less influenced by their connectivity (we performed the experiment with GPT-3.5 turbo agents without a copy mechanism and did not see a difference between dynamic and fully-connected groups). Our empirical observation agrees with previous works that employed groups similar to our random group baseline (Mason et al., 2008; Cantor et al., 2021; Fang et al., 2010), humans (Derex and Boyd, 2016; Migliano et al., 2020) and reinforcement learning agents (Nisioti et al., 2022) and were performed on manually-designed, small knowledge graphs. By generalizing them to a larger knowledge graph grounded in the real world, our work further confirms that social structure matters in innovation tasks and suggests that our testbed can prove useful in future studies of collective innovation with both human and artificial agents.

Discussion

We studied the ability of LLMs in innovation tasks, both in isolation and in groups with different social connectivity. We have shown that GPT-3.5 turbo can leverage the semantics of items to infer the outcomes of crafting but struggle when it comes to planning for multiple time steps and exploring in an open-ended way. Nevertheless, groups of LLMs exhibit an interesting phenomenon previously found in human and computational studies: they perform better collectively when their social connectivity is partial and dynamic rather than static and fully-connected. We attributed this phenomenon to the tree-like structure of the LA2 game: following down some paths may lead you away from other paths and slow down exploration. We have shown that dynamically-connected LLM groups outperform both single LLMs, and fully-connected groups. In groups with dynamic connectivity, subgroups explore different paths and exchange members that share information about other paths, increasing the diversity or breadth of exploration.

Our analysis focused on probing for specific abilities and emergent behaviours in the studied groups of LLMs. However, a larger-scale analysis is necessary to reveal the effect of the different hyperparameters of the model, such as the sampling strategy of LLMs, the configuration of the dynamic connectivity, and the task complexity. A limitation revealed by our empirical study is that smaller, open-source models may fail at learning the task sufficiently well to lead to any interesting emergent behaviours. Thus, as other studies of the cognitive capacities of LLMs have shown, our work suggests that collective innovation studies may require larger models, such as GPT-4 or the introduction of additional mechanisms for complementing their skills. We should note that our experiments did not examine whether pre-training equipped the LLMs with the ability to explore in-context, leverage common-sense knowledge or memorize the solution of Little Alchemy 2. Nevertheless, our conclusion that groups with dynamic connectivity out-compete single-agent and fully-connected ones remains valid.

We believe that this work has important implications. We have shown that groups of LLMs with dynamic connectivity can overcome shortcomings of a single LLM and fully-connected groups. This is a key insight, as dynamic communication structures are less costly than fully-connected ones. LLMs are becoming ubiquitous, participating in various human activities from finance to molecular discovery and writing fiction as copilots of human creativity and productivity (Mirowski et al., 2022; Brinkmann et al., 2023). We believe that this work is a first step towards understanding how and with what kind of connectivity multi-agent LLM systems could optimally and efficiently participate in exploration, innovation, and cultural evolution.

Acknowlegements

This research was partially funded by the French National Research Agency (`https://anr.fr/`, project ECOCURL, Grant ANR-20-CE23-0006) and benefitted from access to the Jean Zay (Idris) supercomputer associated with the Genci grant A0151011996.

References

Acerbi, A. and Stubbersfield, J. M. (2023). Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120. _eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2313790120.

Aplin, L., Farine, D., Morand-Ferron, J., Cockburn, A., Thornton, A., and Sheldon, B. (2015). Experimentally induced innovations lead to persistent culture via conformity in wild birds. *Nature*, 518(7540):538–541.

Aplin, L. M. (2019). Culture and cultural evolution in birds: A review of the evidence. *Animal Behaviour*, 147:179–187.

Boyd, B. (2018). The evolution of stories: from mimesis to language, from fact to fiction. *WIREs Cognitive Science*, 9(1):e1444. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcs.1444.

Boyd, R. and Richerson, P. J. (1996). Why culture is common, but cultural evolution is rare. In *Evolution of Social Behaviour Patterns in Primates and Man*, Proceedings of The British Academy, Vol. 88, pages 77–93. Oxford University Press, New York, NY, US.

Brackbill, D. (2017). The Network Structure Of Collective Innovation. *undefined*.

Brackbill, D. and Centola, D. (2020). Impact of network structure on collective learning: An experimental study in a data science competition. *PLoS ONE*, 15(9):e0237978.

Brinkmann, L., Baumann, F., Bonnefon, J.-F., Derex, M., Müller, T. F., Nussberger, A.-M., Czaplicka, A., Acerbi, A., Griffiths, T. L., Henrich, J., Leibo, J. Z., McElreath, R., Oudeyer, P.-Y., Stray, J., and Rahwan, I. (2023). Machine culture. *Nature Human Behaviour*, 7(11):1855–1868. Number: 11 Publisher: Nature Publishing Group.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. arXiv:2005.14165 [cs].

Brändle, F., Stocks, L. J., Tenenbaum, J. B., Gershman, S. J., and Schulz, E. (2023). Empowerment contributes to exploration behaviour in a creative video game. *Nature Human Behaviour*, 7(9):1481–1489.

Cantor, M., Chimento, M., Smeele, S. Q., He, P., Papageorgiou, D., Aplin, L. M., and Farine, D. R. (2021). Social network architecture and the tempo of cumulative cultural evolution. *Proceedings of the Royal Society B: Biological Sciences*, 288(1946):20203107.

Colas, C., Teodorescu, L., Oudeyer, P.-Y., Yuan, X., and Côté, M.-A. (2023). Augmenting autotelic agents with large language models. In *Conference on Lifelong Learning Agents*, pages 205–226. PMLR.

Creanza, N., Kolodny, O., and Feldman, M. W. (2017). Cultural evolutionary theory: How culture evolves and why it matters. *Proceedings of the National Academy of Sciences*, 114(30):7782–7789. Publisher: Proceedings of the National Academy of Sciences.

Derex, M. (2021). Human cumulative culture and the exploitation of natural phenomena. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1843):20200311.

Derex, M., Beugin, M.-P., Godelle, B., and Raymond, M. (2013). Experimental evidence for the influence of group size on cultural complexity. *Nature*, 503(7476):389–391.

Derex, M. and Boyd, R. (2016). Partial connectivity increases cultural accumulation within groups. *Proceedings of the National Academy of Sciences*, 113(11):2982–2987.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. (2023a). Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv:2305.14325 [cs].

Du, Y., Watkins, O., Wang, Z., Colas, C., Darrell, T., Abbeel, P., Gupta, A., and Andreas, J. (2023b). Guiding Pretraining in Reinforcement Learning with Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 8657–8677. PMLR. ISSN: 2640-3498.

Dunbar, R. I. M. (1993a). Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*, 16(4):681–694.

Dunbar, R. I. M. (1993b). Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*, 16(4):681–694.

Dunbar, R. I. M. (2014). How conversations around campfires came to be. *Proceedings of the National Academy of Sciences*, 111(39):14013.

Eldredge, N. (2011). Paleontology and Cornets: Thoughts on Material Cultural Evolution. *Evolution: Education and Outreach*, 4(3):364–373.

Fang, C., Lee, J., and Schilling, M. (2010). Balancing Exploration and Exploitation Through Structural Design: The Isolation of Subgroups and Organizational Learning. *Organization Science*, 21:625–642.

Fogarty, L., Creanza, N., and Feldman, M. W. (2015). Cultural Evolutionary Perspectives on Creativity and Human Innovation. *Trends in Ecology & Evolution*, 30(12):736–754.

Horton, J. J. (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? arXiv:2301.07543 [econ, q-fin].

Jiang, M., Luketina, J., Nardelli, N., Minervini, P., Torr, P. H. S., Whiteson, S., and Rocktäschel, T. (2020). WordCraft: An Environment for Benchmarking Commonsense Agents. *arXiv:2007.09185 [cs]*.

Kline, M. A. and Boyd, R. (2010). Population size predicts technological complexity in Oceania. *Proceedings of the Royal Society B: Biological Sciences*, 277(1693):2559–2564.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005). All Else Being Equal Be Empowered. In Capcarrère, M. S., Freitas, A. A., Bentley, P. J., Johnson, C. G., and Timmis, J., editors, *Advances in Artificial Life*, pages 744–753, Berlin, Heidelberg. Springer.

Krishnamurthy, A., Harris, K., Foster, D. J., Zhang, C., and Slivkins, A. (2024). Can large language models explore in-context? arXiv:2403.15371 [cs].

Lazer, D. and Friedman, A. (2007). The Network Structure of Exploration and Exploitation. *Administrative Science Quarterly*, 52(4):667–694.

Lotem, A., Halpern, J. Y., Edelman, S., and Kolodny, O. (2017). The evolution of cognitive mechanisms in response to cultural innovations. *Proceedings of the National Academy of Sciences*, 114(30):7915–7922.

Mason, W. and Watts, D. J. (2012). Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, 109(3):764–769.

Mason, W. A., Jones, A., and Goldstone, R. L. (2008). Propagation of innovations in networked groups. *Journal of Experimental Psychology: General*, 137(3):422–433.

Mesoudi, A. (2021). Experimental studies of cultural evolution. preprint, PsyArXiv.

Migliano, A. B., Battiston, F., Viguier, S., Page, A. E., Dyble, M., Schlaepfer, R., Smith, D., Astete, L., Ngales, M., Gomez-Gardenes, J., Latora, V., and Vinicius, L. (2020). Hunter-gatherer multilevel sociality accelerates cumulative cultural evolution. *Science Advances*, 6(9):eaax5913.

Mirowski, P., Mathewson, K. W., Pittman, J., and Evans, R. (2022). Co-Writing Screenplays and Theatre Scripts with Language Models: An Evaluation by Industry Professionals.

Mitchell, M. and Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120. Publisher: Proceedings of the National Academy of Sciences.

Momennejad, I., Hasanbeig, H., Vieira Frujeri, F., Sharma, H., Jojic, N., Palangi, H., Ness, R., and Larson, J. (2023). Evaluating Cognitive Maps and Planning in Large Language Models with CogEval. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 69736–69751. Curran Associates, Inc.

Montuori, A. and Purser, R. E. (1995). Deconstructing the lone genius myth: Toward a contextual view of creativity. *Journal of Humanistic Psychology*, 35(3):69–112. Place: US Publisher: Sage Publications.

Nisioti, E., Mahaut, M., Oudeyer, P.-Y., Momennejad, I., and Moulin-Frier, C. (2022). Social network structure shapes innovation: Experience-sharing in rl with sapiens.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442 [cs].

Pereira, D., Manen, C., and Rigaud, S. (2023). The shaping of social and symbolic capital during the transition to farming in the Western Mediterranean: Archaeological network analyses of pottery decorations and personal ornaments. *PLOS ONE*, 18(11):e0294111. Publisher: Public Library of Science.

Perez, J., Léger, C., Ovando-Tellez, M., Foulon, C., Dussauld, J., Oudeyer, P.-Y., and Moulin-Frier, C. (2024). Cultural evolution in populations of Large Language Models. arXiv:2403.08882 [cs, q-bio].

Perry, S., Carter, A., Smolla, M., Akçay, E., Nöbel, S., Foster, J. G., and Healy, S. D. (2021). Not by transmission alone: the role of invention in cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1828):20200049.

Singh, M., Acerbi, A., Caldwell, C. A., Danchin, É., Isabel, G., Molleman, L., Scott-Phillips, T., Tamariz, M., van den Berg, P., van Leeuwen, E. J. C., and Derex, M. (2021). Beyond social learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1828):20200050.

Smith, D., Schlaepfer, P., Major, K., Dyble, M., Page, A. E., Thompson, J., Chaudhary, N., Salali, G. D., Mace, R., Astete, L., Ngales, M., Vinicius, L., and Migliano, A. B. (2017). Cooperation and the evolution of hunter-gatherer storytelling. *Nature Communications*, 8(1).

Smolla, M., Jansson, F., Lehmann, L., Houkes, W., Weissing, F. J., Hammerstein, P., Dall, S. R. X., Kuijper, B., and Enquist, M. (2021). Underappreciated features of cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1828):20200259. Publisher: Royal Society.

Solé, R. V., Valverde, S., Casals, M. R., Kauffman, S. A., Farmer, D., and Eldredge, N. (2013). The evolutionary ecology of technological innovations. *Complexity*, 18(4):15–27. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cplx.21436.

Stanley, K. O. and Lehman, J. (2015). *Why Greatness Cannot Be Planned: The Myth of the Objective*. Springer International Publishing, Cham.

Tennie, C., Call, J., and Tomasello, M. (2009). Ratcheting up the ratchet: On the evolution of cumulative culture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1528):2405–2415.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. (2023a). Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. (2023b). Voyager: An Open-Ended Embodied Agent with Large Language Models. arXiv:2305.16291 [cs].

Webb, T., Mondal, S. S., Wang, C., Krabach, B., and Momennejad, I. (2024). A Prefrontal Cortex-inspired Architecture for Planning in Large Language Models. arXiv:2310.00194 [cs].

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Xia, F., Le, Q., and Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. *ArXiv*.

Whiten, A., Biro, D., Bredeche, N., Garland, E. C., and Kirby, S. (2021a). The emergence of collective knowledge and cumulative culture in animals, humans and machines. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1843):20200306. Publisher: Royal Society.

Whiten, A., Biro, D., Bredeche, N., Garland, E. C., and Kirby, S. (2021b). The emergence of collective knowledge and cumulative culture in animals, humans and machines. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1843):20200306.

Whiten, A., Goodall, J., McGrew, W. C., Nishida, T., Reynolds, V., Sugiyama, Y., Tutin, C. E. G., Wrangham, R. W., and Boesch, C. (1999). Cultures in chimpanzees. *Nature*, 399(6737):682–685.

Zhuge, M., Liu, H., Faccio, F., Ashley, D. R., Csordás, R., Gopalakrishnan, A., Hamdi, A., Hammoud, H. A. A. K., Herrmann, V., Irie, K., et al. (2023). Mindstorms in natural language-based societies of mind. *arXiv preprint arXiv:2305.17066*.